# Mining mass spectrometry data: Using new computational tools to find novel organic compounds in complex environmental mixtures

Krista Longnecker* and Elizabeth B. Kujawinski

Woods Hole Oceanographic Institution, Department of Marine Chemistry and Geochemistry, Woods Hole, MA, 02543 USA

*Corresponding author: klongnecker@whoi.edu

Short title: Lyso-sulfolipids in marine samples

Accepted by *Organic Geochemistry*

Keywords : lyso-sulfolipids; sulfoquinovosyl head group; metabolomics; fragmentation spectra; molecular networking

**Abstract**

Untargeted metabolomics datasets provide ample opportunity for discovery of novel metabolites. The major challenge is focusing data analysis on a short list of metabolites. Here, we apply a combination of computational tools that serve to reduce complex mass spectrometry data in order allow us to focus on new environmentally-relevant metabolites. In the first portion of the project, we explored mass spectrometry data from intracellular metabolites extracted from a model marine diatom, *Thalassiosira pseudonana*. The fragmentation data from these samples were analyzed using molecular networking, an on-line tool that clusters metabolites based on shared structural similarities. The features within each metabolite cluster were then putatively annotated using MetFrag, an *in silico* fragmentation tool. Using this combination of computational tools, we observed multiple lyso-sulfolipids, organic compounds not previously known to exist within cultured marine diatoms. In the second stage of the project, we searched our environmental data for these lyso-sulfolipids. The lyso-sulfolipid with a C14:0 fatty acid was found in dissolved and particulate samples from the western Atlantic Ocean, and a culture of cyanobacteria grown in our laboratory. Thus, the putative lyso-sulfolipids are present in both laboratory experiments and environmental samples. This project highlights the value of combining computational tools to detect and putatively identify organic compounds not previously recognized as important within *T. pseudonana* or the marine environment. Future applications of these tools to emerging metabolomics data will further open the black box of natural organic matter, identifying molecules that can be used to understand and monitor the global carbon cycle.

## Introduction

Organic matter is a complex and heterogeneous mixture of compounds that challenges scientists investigating its role in global biogeochemical processes. Organic compounds are formed from inorganic carbon through the actions of primary producers. These compounds can then be transformed into new organic compounds through biological activity, and may ultimately be converted back to inorganic carbon. Through these actions, the biological processes and the composition of organic matter are tightly coupled (Azam et al., 1993). Each organic compound also has its own source and sink dynamic, which potentially varies with biotic and abiotic parameters in an ecosystem. Yet, we have only identified a small fraction of organic compounds that exist in the environment and have limited understanding of the roles of these compounds in the carbon cycle. Organic matter likely contains thousands of individual molecules, making comprehensive identification an elusive goal. However, biologically-derived molecules such as metabolites are likely to play an important role in the carbon cycle, either as growth substrates or growth factors for microbes. Thus identification of these molecules could provide insights into the function and metabolism of microbes that govern the ocean carbon cycle (Moran et al., 2016).

We present a novel combination of computational tools with the goal of more efficiently identifying individual compounds within a complex mixture of organic matter. This project expands our ability to analyze untargeted metabolomics data and is one of several methods that can be used to characterize organic matter in aquatic environments (e.g., Longnecker et al.,

2015a; Longnecker and Kujawinski, 2016; Treutler et al., 2016; van der Hooft et al., 2016). Here, our analysis is based on two modes of analyzing fragmentation spectra from organic molecules. These fragmentation spectra can be grouped based on the similarity of fragment *m*/*z* values measured within a set of samples (Frank et al., 2007; Nguyen et al., 2013). This clustering of fragmentation spectra, also called molecular networking (Yang et al., 2013), has proven useful in finding known compounds within microbial colonies growing in the laboratory (Watrous et al., 2012) and in environmental samples (Kharbush et al., 2016; Teta et al., 2015). Here, we combine molecular networking with MetFrag, an *in silico* fragmentation tool (Wolf et al., 2010) that presents potential compound identifications given a measured fragmentation spectra.

We introduce our approach through a comparison of laboratory data and environmental data. The laboratory data were intracellular metabolites extracted from the centric diatom *Thalassiosira pseudonana* which was grown under phosphate-limited and phosphate-replete conditions. Thousands of intracellular metabolites are produced by *T. pseudonana*, yet our previous research revealed that most of these metabolites cannot be identified (Longnecker et al., 2015b). Here, we were able to identify a set of metabolites not previously known to be important within diatom physiology and then expanded our analysis to investigate the extent to which these compounds were found within marine ecosystems.

## Materials and Methods

### Untargeted metabolomics experiments with a cultured marine diatom

*Thalassiosira pseudonana* (CCMP #1335) was cultured axenically in modified L1 media. There were two treatments: phosphate-replete (36 μM $PO_4^{-3}$) and phosphate-limited (0.4 μM $PO_4^{-3}$). The experiment began with the addition of 30 ml of *T. pseudonana* in exponential phase to two-thirds of the flasks which contained 300 ml of media. The remaining one-third of the flasks were designated as cell-free controls. The cultures were maintained under a 12:12 light:dark regime (84 μmol $m^{-2}$ $s^{-1}$). Two flasks with cells and one cell-free control for each treatment were sampled at four time points: 0, 2, 8, and 10 days. At each time point, cells were captured by gentle vacuum filtration on 0.2 μm Omnipore (Millipore) filters. The intracellular metabolites were extracted using a method modified from Rabinowitz and Kimball (2007), as described previously (Kido Soule et al., 2015). The extracts were re-dissolved in 95:5 water:acetonitrile and deuterated biotin (final concentration 0.05 μg $ml^{-1}$) and analyzed in negative ion mode with liquid chromatography (LC) coupled by electrospray ionization to a 7-Tesla Fourier-transform ion cyclotron resonance mass spectrometer (Thermo Scientific, FT-ICR MS). LC separation was performed using a Synergi Fusion reversed phase column (Phenomenex, Torrance, CA). The chromatography gradient was: an initial hold of 95% A (0.1% formic acid in water) : 5% B (0.1% formic acid in acetonitrile) for 2 minutes, ramp to 65% B from 2 to 20 minutes, ramp to 100% B from 20 to 25 min, and hold until 32.5 minutes. The column was re-equilibrated for 7 min between samples with solvent A. In parallel to the FT acquisition, four data-dependent fragmentation (MS/MS) scans were collected at nominal mass resolution in the ion trap (LTQ).

Samples were analyzed in random order with a pooled sampled run every six samples in order to assess instrument variability.

The data files from the mass spectrometer were converted to the open-source mzXML format using the MSConverter tool (Kessner et al., 2008). After this step, the data files were processed in parallel using two different data analysis pipelines (Figure 1). The first pipeline involves the use of XCMS (Smith et al., 2006) to conduct the peak picking, alignment, and retention time correction that is standard for untargeted metabolomics data analysis (Johnson et al., 2014). The output from this analysis is a list of 'mzRT features' which are defined as unique combinations of *m*/*z* values and retention times that have passed our quality control checks. The peak area for each mzRT feature provides an estimate of the relative levels of the feature during the experiment. The freely-available XCMS2 code (Benton et al., 2008) was used to generate the list of MS2 spectra obtained for the mzRT features and CAMERA provided details about possible isotopologues within the dataset (Kuhl et al., 2012). The untargeted metabolomics data from the *T. pseudonana* cultures are available with accession code MTBLS154 at MetaboLights (Haug et al., 2013).

## Molecular networking

Molecular networking is one of the functions available at the Global Natural Products Social Molecular Networking site (GNPS, http://gnps.ucsd.edu) that has recently been described by Wang et al. (2016). We used the molecular networking tool (Yang et al., 2013) to group our mzRT features based on similarities in the MS2 fragmentation spectra. Molecular networking

takes advantage of the MS2 information within unprocessed mzXML files. The *T. pseudonana* data used here are available at GNPS with MassIVE ID MSV000080990. The molecular networking analysis was run with the following parameters: parent mass tolerance = 1 Da, ion tolerance = 0.5 Da, minimum cluster size = 3, minimum pairs cosine = 0.7, score threshold = 0.5, network topK = 10, run MSCluster = TRUE. The output is a network calculated based on the overlap in peaks within the MS2 spectra. Within the network, each node is an mzRT feature with MS2 data. The nodes are connected by edges, the width of each edge is a measure of the similarity in the MS2 spectra between two nodes. We used Cytoscape (Smoot et al., 2011) as a visualization tool to annotate the nodes to include information about each mzRT feature (experimental treatment, *m*/*z* value, and strength of connection with other mzRT features within the cluster).

## *In silico* calculation of fragmentation spectra

We used MetFrag (Wolf et al., 2010) to putatively annotate mzRT features based on their MS2 spectra. The analysis starts with a search for exact mass of the parent ion against a database; the available database options are currently KEGG, PubChem, and ChemSpider. Once potential matches are located, MetFrag generates *in silico* fragments from the parent compound and compares the *in silico* fragments to the measured MS2 fragments uploaded by the user. We used the data output from the XCMS processing (left side of Figure 1) to provide the *m*/*z* values and peak intensities that were used in the MetFrag search. This required a computational step that matched the *m*/*z* values and retention times from the XCMS output with the results file

produced by the molecular networking tool. While this required extra steps, the protocol allowed us to use the data that passed our quality control checks within XCMS in lieu of relying on the unprocessed mzXML files used by the molecular networking tool. The fragmentation spectra from the mzRT features of interest were searched in MetFrag using a 1 ppm window for the parent ion search in the ChemSpider database. For the fragments we set MZabs = 1, and MZppm = 30. The output was manually inspected to assess possible annotations.

## Comparisons to existing untargeted metabolomics data

We used the domdb database (Longnecker et al., 2015a) to compare select metabolites from the *T. pseudonana* cultures with our existing untargeted metabolomics data. The database allows searching by *m*/*z* and retention time. The search thus requires samples to be analyzed using the same analytical methods to minimize variability in retention times that occur with different liquid chromatography conditions. One of the metabolites discussed below was found in four sets of samples processed within our laboratory: a culture experiment with *Synechococcus elongatus*, incubation experiments with Atlantic Ocean seawater collected from 70 m and 700 m, and sinking particles collected from net traps deployed for 24 hours at 150 m at select stations in the Atlantic Ocean (Table 2).

## Results and Discussion

In the following sections, we start with untargeted metabolomics data from a laboratory experiment to reveal the value of integrated molecular networking and *in silico* analysis of fragmentation spectra. Individually, each of these tools provides valuable insights into mass

spectrometry data. Collectively, the combination allows for efficient data mining into the putative annotations of select metabolites. The time needed to putatively identify metabolites is the most time consuming aspect of metabolomics experiments. Thus technical advances, such our novel combination of computational tools, provide a new means for researchers to focus on ecologically interesting results.

## Molecular networking: clustering of mzRT features by MS2 spectra

Molecular networking formed clusters within the *T. pseudonana* data based on the similarity between the MS2 spectra (Figure 2). With the parameters we selected at GNPS, we obtained a set of thirteen clusters representing 90 mzRT features within our dataset. In comparison, 2825 mzRT features were in the final, aligned dataset from the XCMS analysis and 1303 of these mzRT features had associated MS2 spectra. Thus, molecular networking helped to constrain the dataset to a tractable number of features for additional analysis. Alternatively, a user could set the input parameters at GNPS to allow a more conservative analysis with higher numbers of linked mzRT features. In our analysis, the clusters linked a minimum of two mzRT features (e.g, clusters K, L, or M) and maximum of 19 mzRT features (cluster A). The different colors in the network diagram indicate whether an mzRT feature is present only in the phosphate-replete or phosphate-limited conditions, or is present under both growth conditions. For example, the cluster marked A contains mzRT features present under both growth conditions individually as well as together, while cluster I only contains mzRT features present during phosphate-limited growth. These mzRT features represent promising targets for

markers that are unique to specific ecological conditions. Yet, because molecular networking considers unprocessed mzXML files, there are poor-quality mzRT features present as nodes in the network (i.e., the mzRT features that do not pass the quality control checks in the XCMS processing, see Methods). For example, all the mzRT features in cluster F were removed by XCMS processing.

One benefit of the molecular networking tool is its ability to group a compound with its isotopologues. The MS2 fragments from a compound and its isotopologues will have similar fragmentation spectra, even though the $m/z$ values for the fragments may differ by the mass difference between $^{12}C$ and $^{13}C$. In cluster A, six of the mzRT features are paired sets of metabolites with the charged compound as one node and the charged compound with a single $^{13}C$ atom as a second node within the cluster. The molecular networking tool does not annotate these MS2 fragmentation spectra as originating from a compound and its isotopologues; rather this distinction was identified by the CAMERA algorithm. In order to identify isotopologues within the clusters, we combined the output from the molecular networking results with the processed XCMS/CAMERA results. The integration of these two outputs streamlined our identification efforts by removing $^{13}C$ compounds from further analysis with $^{12}C$-based computational tools such as MetFrag.

The GNPS website also provides users with the opportunity to compare measured fragmentation spectra with fragmentation spectra stored at GNPS (Wang et al., 2016). The fragmentation spectra stored at GNPS originate from any user consenting to the public use of

193 their data. However, in the case of the *T. pseudonana* dataset, none of the mzRT features had a
194 corresponding match to a metabolite in the GNPS database. Thus, while GNPS contains
195 increasing numbers of fragmentation spectra, the database is not yet a comprehensive survey of
196 organic compounds from environmental mixtures. Yet, even without the database match,
197 inspection of the nodes within cluster A revealed a set of mzRT features with direct relevance to
198 diatom physiology in marine environments, which could be putatively identified with MetFrag.

199 **Putatively annotating mzRT features within a cluster based on fragmentation spectra**

200 The identification of unknown metabolites is often a primary goal of untargeted
201 metabolomics projects as researchers seek to quantify the biogeochemical cycling of known
202 organic compounds. We used the classification scheme defined by Sumner et al. (2007) to guide
203 our descriptions of the putative metabolite annotations. Within this scheme, the metabolites we
204 discuss below are Level 2 identifications which are putatively annotated without chemical
205 reference standards, but are based on spectral similarities with data from public or commercial
206 libraries.

207 The similarities in the MS2 spectra in the mzRT features grouped by molecular
208 networking into cluster A is evident when the MS2 fragments are plotted together (Figure 3).
209 Note that all four of the mzRT features plotted in Figure 3 were observed as the charged ion and
210 the isotopologue with a single $^{13}C$ atom; Figure S1 shows the corresponding plots of the MS2
211 fragments from $^{13}C$ compounds. All four of these mzRT features have a sulfoquinovosyl head
212 group (sulfoquinovose, Figure 3A), which is a derivative of glucose with the 6-hydroxyl

replaced by a sulfonate group (Benning, 1998). Three of the mzRT features differ by the fatty acid chain, with 14:0, 16:0, and 16:1 as potential options. The 14 and 16 refer to the number of carbon atoms in the fatty acid while the 0 or 1 refers to the number of double bonds within the fatty acid. Figure S2 shows the structure of each of these metabolites and Table S1 includes images of the MS2 fragments and the distribution of the fragments across the mzRT features from *T. pseudonana*. As further support of our putative annotation of these mzRT features, three of the fragments we measured were noted as characteristic fragments of sulfoquinovosyl monoacylglycerols (SQMG) by De Souza et al. (2006). More generally, these SQMG compounds are known as lyso-sulfolipids. While analysis of these mzRT features with authentic standards would be ideal, these compounds are not commercially available. Thus, in the absence of such standards, the putative annotation of these compounds is state-of-the art.

## Lyso-sulfolipids in diatoms

Lipids are the structural underpinning of the bilayer membrane surrounding a cell. However, the lipids that comprise cell membranes have a polar head group and two non-polar fatty acid tails, and this combination causes the lipids to self-assemble into a bilayer membrane. In contrast, the lyso-sulfolipids observed in the present project have only a single fatty acid. The biochemical origin of these lyso-sulfolipids is unknown, but here we consider several possibilities. The lyso-sulfolipids could have been derived from sulfoquinovosyl diacylglycerols (SQDG), the corresponding sulfolipid with two fatty acids which is an essential component of photosynthetic membranes. This process has been observed to be enzymatically possible in

some (Gupta and Sastry, 1987; Wolfersberger and Pieringer, 1974; Yagi and Benson, 1962), but not all organisms (Burns et al., 1977). Alternatively, only a single fatty acid could be combined with the sulfoquinovosyl head group to form the lyso-sulfolipid. Finally, the lipids could have degraded during sample processing (Allen et al., 1970), although we consider this option less likely because the filters were stored frozen at -80° C and analyzed using mass spectrometry within 10 days after extraction. Additional research will be needed to determine which process is occurring within our samples.

Lyso-sulfolipids have been observed in cultures of marine algae (El Baz et al., 2013), but not, to our knowledge, within cultures of *T. pseudonana*. Yet, SQDG lipids play a prominent role in *T. pseudonana*'s physiological response to phosphorus limitation as the diatom switches from phosphorus based lipids to sulfolipids in order to spare phosphorus for other cellular functions (Martin et al., 2011; Van Mooy et al., 2009). SQDG has also been observed in single cell measurements of *Chlamydomonas* grown under nitrogen limited conditions (Cahill et al., 2015), which may indicate a broad physiological need for SQDG under nutrient limited growth. The ecological role of lyso-sulfolipids within the metabolism of *T. pseudonana* is not known. Yet, as with SQDG, these lipids are more prevalent under conditions of phosphate-limited growth (Figure 4), which suggests the lyso-sulfolipids are also playing a role in phosphorus scavenging within the cells.

In addition to the lyso-sulfolipids, we also putatively annotated the sulfoquinovosyl head group attached to a 21-carbon sterol (Figure 5). On a per-cell basis, this metabolite was more 1.7

253 times more prevalent under phosphate-limited growth conditions, although a significant
254 amount of the compound was also found in the phosphate-replete cultures. *T. pseudonana*, like
255 all eukaryotes, makes sterol compounds and uses them to maintain the structural integrity of its
256 cell membrane. Sterols in *T. pseudonana* are primarily 27- or 28-carbon sterols (Rampen et al.,
257 2010; Véron et al., 1998), larger than the 21-carbon sterol we observed. In the marine
258 environment, 21-carbon sterols are not common, although they are present in marine sponges
259 (Ballantine et al., 1977). Given the novelty of a sulfoquinovosyl head group attached to a sterol,
260 we cannot speculate as to the role of this compound within the metabolism of *T. pseudonana*.

261 **Additional observations of lyso-sulfolipids in marine samples**

262 The lyso-sulfolipids are not unique to laboratory cultures with *T. pseudonana*. Using our
263 domdb database, we found the C14:0 lyso-sulfolipid in four sets of samples processed by our
264 laboratory using the same methods described for *T. pseudonana* (Table 2). In three cases, we have
265 extracts from paired filters and filtrate samples (Figure 6A, B, and C); the C14:0 lyso-sulfolipid
266 was always found at elevated levels in the filtrate compared to the filters. The C14:0 lyso-
267 sulfolipid was also found in particulate material captured by net traps deployed for 24 hours at
268 150 m (Figure 6D); no filtrate was processed for the net trap samples. None of the other
269 compounds listed in Table 1 were found in any of our samples, nor was the C14:0 lyso-
270 sulfolipid found in the filtrate from the experiment with *T. pseudonana* or in any of our sample
271 processing or instrumentation blanks (data not shown). Yet, the presence of the C14:0 lyso-
272 sulfolipid in samples spanning from the surface ocean, to deep seawater, and to laboratory

cultures hints at the prospect of a set of organic compounds that may provide information on the physiological state of organisms in marine environments.

## Conclusions and ecological significance

We used a combination of molecular networking and *in silico* fragmentation computational tools to find a novel class of lipids within a set of ultrahigh resolution mass spectrometry data. Without this combination of computational tools, we would not have focused on putatively identifying these compounds, nor would we have known to look within our existing data to find other sources for lyso-sulfolipids. The lyso-sulfolipids were the only compounds to result from the computational tools described here. Beyond their classification as lyso-sulfolipids, the compounds described here are sulfur-containing organic molecules which are increasingly recognized as important within marine ecosystems (Ksionzek et al., 2016). Organic sulfur compounds are transferred from autotrophic to heterotrophic microorganisms (Durham et al., 2015; Malmstrom et al., 2004). In the process, select organic sulfur compounds serve as signaling molecules to which heterotrophic bacteria respond (Johnson et al., 2016). Lyso-sulfolipids have been shown to serve as signaling molecules and induce larval settlement and metamorphosis in sea urchins (Takahashi et al., 2002) and corals (Tebben et al., 2015). Finally, while the sulfoquinovosyl head group can be degraded to other organic sulfur compounds (Denger et al., 2012; Felux et al., 2015), we did not observe any of the degradation products within the particulate material sampled during these experiments (data not shown). Given the presence of the lyso-sulfolipids in multiple experiments and field samples, we posit

293 that these compounds are serving an active role within the physiology of microbial cells.
294 Furthermore, the presence of the C14:0 lyso-sulfolipid in water samples from the Atlantic Ocean
295 provides a direct link between compounds found in laboratory cultures and compounds
296 observed in the marine environment.

297 **Acknowledgements**

298 We thank Crystal Breier, Gretchen Swarr, and Bill Arnold for assistance in the laboratory,
299 and the captain and crew of the R/V *Knorr* and the DeepDOM cruise participants for assistance
300 obtaining the Atlantic Ocean samples. We are especially appreciative of the work done by
301 Benjamin Van Mooy and Justin Ossolinski that allowed us to obtain particulate material from
302 their net traps. Discussions with Melissa Kido Soule, Helen Fredricks, Sean Sylva, and Jamey
303 Fulton about the composition of the organic compounds were enlightening. Collection of the
304 field samples was funded by NSF OCE-1154320 (to EBK and KL). The culture experiments and
305 subsequent data analysis were funded by the Gordon and Betty Moore Foundation through
306 Grant GBMF3304 to EBK.

## References


Allen, C.F., Good, P., Holton, R.W., 1970. Lipid composition of *Cyanidium*. Plant Physiology 46, 748-751.

Azam, F., Smith, D.C., Steward, G.F., Hagström, Å., 1993. Bacteria-organic matter coupling and its significance for oceanic carbon cycling. Microbial Ecology 28, 167-179.

Ballantine, J.A., Williams, K., Burke, B.A., 1977. Marine sterols IV. C21 sterols from marine sources. Identification of pregnane derivatives in extracts of the sponge *Haliclona rubens*. Tetrahedron Letters 18, 1547-1550.

Benning, C., 1998. Biosynthesis and function of the sulfolipid sulfoquinovosyl diacylglycerol. Annual Review of Plant Physiology and Plant Molecular Biology 49, 53-75.

Benton, H.P., Wong, D.M., Trauger, S.A., Siuzdak, G., 2008. XCMS$^2$: Processing tandem mass spectrometry data for metabolite identification and structural characterization. Analytical Chemistry 80, 6382-6389.

Burns, D.D., Galliard, T., Harwood, J.L., 1977. Catabolism of sulfoquinovosyl diacylglycerol by an enzyme preparation from *Phaseolus multiflorus*. Phytochemistry 16, 651-654.

Cahill, J.F., Darlington, T.K., Fitzgerald, C., Schoepp, N.G., Beld, J., Burkart, M.D., Prather, K.A., 2015. Online analysis of single cyanobacteria and algae cells under nitrogen-limited conditions using aerosol time-of-flight mass spectrometry. Analytical Chemistry 87, 8039-8046.

De Souza, D.P., Saunders, E.C., McConville, M.J., Likic, V.A., 2006. Progressive peak clustering in GC-MS metabolomic experiments applied to *Leishmania* parasites. Bioinformatics 22, 1391 - 1396.

Denger, K., Huhn, T., Hollemeyer, K., Schleheck, D., Cook, A.M., 2012. Sulfoquinovose degraded by pure cultures of bacteria with release of $C_3$-organosulfonates: complete degradation in two-member communities. FEMS Microbiology Letters 328, 39-45.

Durham, B.P., Sharma, S., Luo, H., Smith, C.B., Amin, S.A., Bender, S.J., Dearth, S.P., Van Mooy, B.A.S., Campagna, S.R., Kujawinski, E.B., Armbrust, E.V., Moran, M.A., 2015. Cryptic carbon and sulfur cycling between surface ocean plankton. Proceedings of the National Academy of Sciences 112, 453-457.

El Baz, F.K., Baroty, G.S.E., Baky, H.H.A.E., El-Salam, O.I.A., Ibrahim, E.A., 2013. Structural characterization and biological activity of sulfolipids from selected marine algae. Grasas y Aceites 64, 561-571.

Felux, A.-K., Spiteller, D., Klebensberger, J., Schleheck, D., 2015. Entner–Doudoroff pathway for sulfoquinovose degradation in *Pseudomonas putida* SQ1. Proceedings of the National Academy of Sciences 112, E4298-E4305.

Fiore, C.L., Longnecker, K., Kido Soule, M.C., Kujawinski, E.B., 2015. Release of ecologically relevant metabolites by the cyanobacterium, *Synechococcus elongatus* CCMP 1631. Environmental Microbiology 17, 3949-3963.

Frank, A.M., Bandeira, N., Shen, Z., Tanner, S., Briggs, S.P., Smith, R.D., Pevzner, P.A., 2007. Clustering millions of tandem mass spectra. Journal of Proteome Research 7, 113-122.

Gupta, S.D., Sastry, P.S., 1987. Metabolism of the plant sulfolipid-Sulfoquinovosyldiacylglycerol: Degradation in animal tissues. Archives of Biochemistry and Biophysics 259, 510-519.

Haug, K., Salek, R.M., Conesa, P., Hastings, J., de Matos, P., Rijnbeek, M., Mahendraker, T., Williams, M., Neumann, S., Rocca-Serra, P., Maguire, E., González-Beltrán, A., Sansone, S.-A., Griffin, J.L., Steinbeck, C., 2013. MetaboLights—an open-access general-purpose repository for metabolomics studies and associated meta-data. Nucleic Acids Research 41, D781-D786.

Johnson, C.H., Ivanisevic, J., Benton, H.P., Siuzdak, G., 2014. Bioinformatics: the next frontier of metabolomics. Analytical Chemistry 87, 147-156.

Johnson, W.M., Kido Soule, M.C., Kujawinski, E.B., 2016. Evidence for quorum sensing and differential metabolite production by a marine bacterium in response to DMSP. ISME Journal.

Kessner, D., Chambers, M., Burke, R., Agus, D., Mallick, P., 2008. ProteoWizard: open source software for rapid proteomics tools development. Bioinformatics 24, 2534-2536.

Kharbush, J.J., Allen, A.E., Moustafa, A., Dorrestein, P.C., Aluwihare, L.I., 2016. Intact polar diacylglycerol biomarker lipids isolated from suspended particulate organic matter accumulating in an ultraoligotrophic water column. Organic Geochemistry 100, 29-41.

Kido Soule, M.C., Longnecker, K., Johnson, W.M., Kujawinski, E.B., 2015. Environmental metabolomics: analytical strategies. Marine Chemistry 177, Part 2, 374-387.

Ksionzek, K.B., Lechtenfeld, O.J., McCallister, S.L., Schmitt-Kopplin, P., Geuer, J.K., Geibert, W., Koch, B.P., 2016. Dissolved organic sulfur in the ocean: Biogeochemistry of a petagram inventory. Science 354, 456-459.

Kuhl, C., Tautenhahn, R., Böttcher, C., Larson, T.R., Neumann, S., 2012. CAMERA: an integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets. Analytical Chemistry 84, 283-289.

Longnecker, K., Futrelle, J., Coburn, E., Kido Soule, M.C., Kujawinski, E.B., 2015a. Environmental metabolomics: databases and tools for data analysis. Marine Chemistry 177, Part 2, 366-373.

Longnecker, K., Kido Soule, M.C., Kujawinski, E.B., 2015b. Dissolved organic matter produced by *Thalassiosira pseudonana*. Marine Chemistry 168, 114-123.

Longnecker, K., Kujawinski, E.B., 2016. Using network analysis to discern compositional patterns in ultrahigh resolution mass spectrometry data of dissolved organic matter. Rapid Communications in Mass Spectrometry 30, 2388-2394.

Malmstrom, R.R., Kiene, R.P., Kirchman, D.L., 2004. Identification and enumeration of bacteria assimilating dimethylsulfoniopropionate (DMSP) in the North Atlantic and Gulf of Mexico. Limnology and Oceanography 49, 597-606.

Martin, P., Van Mooy, B.A.S., Heithoff, A., Dyhrman, S.T., 2011. Phosphorus supply drives rapid turnover of membrane phospholipids in the diatom *Thalassiosira pseudonana*. ISME Journal 5, 1057-1060.

Moran, M.A., Kujawinski, E.B., Stubbins, A., Fatland, R., Aluwihare, L.I., Buchan, A., Crump, B.C., Dorrestein, P.C., Dyhrman, S.T., Hess, N.J., Howe, B., Longnecker, K., Medeiros, P.M., Niggemann, J., Obernosterer, I., Repeta, D.J., Waldbauer, J.R., 2016. Deciphering ocean carbon in a changing world. Proceedings of the National Academy of Sciences 113, 3143-3151.

Nguyen, D.D., Wu, C.-H., Moree, W.J., Lamsa, A., Medema, M.H., Zhao, X., Gavilan, R.G., Aparicio, M., Atencio, L., Jackson, C., Ballesteros, J., Sanchez, J., Watrous, J.D., Phelan, V.V., van de Wiel, C., Kersten, R.D., Mehnaz, S., De Mot, R., Shank, E.A., Charusanti, P., Nagarajan, H., Duggan, B.M., Moore, B.S., Bandeira, N., Palsson, B.Ø., Pogliano, K., Gutiérrez, M., Dorrestein, P.C., 2013. MS/MS networking guided analysis of molecule and gene cluster families. Proceedings of the National Academy of Sciences 110, E2611-E2620.

Rabinowitz, J.D., Kimball, E., 2007. Acidic acetonitrile for cellular metabolome extraction from *Escherichia coli*. Analytical Chemistry 79, 6167-6173.

Rampen, S.W., Abbas, B.A., Schouten, S., Damste, J.S.S., 2010. A comprehensive study of sterols in marine diatoms (Bacillariophyta): Implications for their use as tracers for diatom productivity. Limnology and Oceanography 55, 91-105.

Smith, C.A., Want, E.J., O'Maille, G., Abagyan, R., Siuzdak, G., 2006. XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. Analytical Chemistry 78, 779 - 787.

402 Smoot, M.E., Ono, K., Ruscheinski, J., Wang, P.-L., Ideker, T., 2011. Cytoscape 2.8: new features
403 for data integration and network visualization. Bioinformatics 27, 431-432.

404 Sumner, L., Amberg, A., Barrett, D., Beale, M., Beger, R., Daykin, C., Fan, T., Fiehn, O.,
405 Goodacre, R., Griffin, J., Hankemeier, T., Hardy, N., Harnly, J., Higashi, R., Kopka, J., Lane, A.,
406 Lindon, J., Marriott, P., Nicholls, A., Reily, M., Thaden, J., Viant, M., 2007. Proposed minimum
407 reporting standards for chemical analysis. Metabolomics 3, 211-221.

408 Takahashi, Y., Itoh, K., Ishii, M., Suzuki, M., Itabashi, Y., 2002. Induction of larval settlement
409 and metamorphosis of the sea urchin Strongylocentrotus intermedius by glycoglycerolipids
410 from the green alga Ulvella lens. Marine Biology 140, 763-771.

411 Tebben, J., Motti, C.A., Siboni, N., Tapiolas, D.M., Negri, A.P., Schupp, P.J., Kitamura, M., Hatta,
412 M., Steinberg, P.D., Harder, T., 2015. Chemical mediation of coral larval settlement by crustose
413 coralline algae. Scientific Reports 5.

414 Teta, R., Della Sala, G., Glukhov, E., Gerwick, L., Gerwick, W.H., Mangoni, A., Costantino, V.,
415 2015. Combined LC–MS/MS and molecular networking approach reveals new cyanotoxins from
416 the 2014 cyanobacterial bloom in Green Lake, Seattle. Environmental Science & Technology 49,
417 14301-14310.

418 Treutler, H., Tsugawa, H., Porzel, A., Gorzolka, K., Tissier, A., Neumann, S., Balcke, G.U., 2016.
419 Discovering Regulated Metabolite Families in Untargeted Metabolomics Studies. Analytical
420 Chemistry 88, 8082-8090.

421 van der Hooft, J.J.J., Wandy, J., Barrett, M.P., Burgess, K.E.V., Rogers, S., 2016. Topic modeling
422 for untargeted substructure exploration in metabolomics. Proceedings of the National Academy
423 of Sciences 113, 13738-13743.

424 Van Mooy, B.A.S., Fredricks, H.F., Pedler, B.E., Dyhrman, S.T., Karl, D.M., Koblížek, M., Lomas,
425 M.W., Mincer, T.J., Moore, L.R., Moutin, T., Rappé, M.S., Webb, E.A., 2009. Phytoplankton in the
426 ocean use non-phosphorus lipids in response to phosphorus scarcity. Nature 458, 69-72.

427 Véron, B., Dauguet, J.-C., Billard, C., 1998. Sterolic biomarkers in marine phytoplankton. II. Free
428 and conjugated sterols of seven species used in mariculture. Journal of Phycology 34, 273-279.

429 Wang, M., Carver, J.J., Phelan, V.V., Sanchez, L.M., Garg, N., Peng, Y., Nguyen, D.D., Watrous,
430 J., Kapono, C.A., Luzzatto-Knaan, T., Porto, C., Bouslimani, A., Melnik, A.V., Meehan, M.J., Liu,
431 W.-T., Crusemann, M., Boudreau, P.D., Esquenazi, E., Sandoval-Calderon, M., Kersten, R.D.,
432 Pace, L.A., Quinn, R.A., Duncan, K.R., Hsu, C.-C., Floros, D.J., Gavilan, R.G., Kleigrewe, K.,
433 Northen, T., Dutton, R.J., Parrot, D., Carlson, E.E., Aigle, B., Michelsen, C.F., Jelsbak, L.,
434 Sohlenkamp, C., Pevzner, P., Edlund, A., McLean, J., Piel, J., Murphy, B.T., Gerwick, L., Liaw,
435 C.-C., Yang, Y.-L., Humpf, H.-U., Maansson, M., Keyzers, R.A., Sims, A.C., Johnson, A.R.,

Sidebottom, A.M., Sedio, B.E., Klitgaard, A., Larson, C.B., Boya P, C.A., Torres-Mendoza, D., Gonzalez, D.J., Silva, D.B., Marques, L.M., Demarque, D.P., Pociute, E., O'Neill, E.C., Briand, E., Helfrich, E.J.N., Granatosky, E.A., Glukhov, E., Ryffel, F., Houson, H., Mohimani, H., Kharbush, J.J., Zeng, Y., Vorholt, J.A., Kurita, K.L., Charusanti, P., McPhail, K.L., Nielsen, K.F., Vuong, L., Elfeki, M., Traxler, M.F., Engene, N., Koyama, N., Vining, O.B., Baric, R., Silva, R.R., Mascuch, S.J., Tomasi, S., Jenkins, S., Macherla, V., Hoffman, T., Agarwal, V., Williams, P.G., Dai, J., Neupane, R., Gurr, J., Rodriguez, A.M.C., Lamsa, A., Zhang, C., Dorrestein, K., Duggan, B.M., Almaliti, J., Allard, P.-M., Phapale, P., Nothias, L.-F., Alexandrov, T., Litaudon, M., Wolfender, J.-L., Kyle, J.E., Metz, T.O., Peryea, T., Nguyen, D.-T., VanLeer, D., Shinn, P., Jadhav, A., Muller, R., Waters, K.M., Shi, W., Liu, X., Zhang, L., Knight, R., Jensen, P.R., Palsson, B.O., Pogliano, K., Linington, R.G., Gutierrez, M., Lopes, N.P., Gerwick, W.H., Moore, B.S., Dorrestein, P.C., Bandeira, N., 2016. Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. Nature Biotechnology 34, 828-837.

Watrous, J., Roach, P., Alexandrov, T., Heath, B.S., Yang, J.Y., Kersten, R.D., van der Voort, M., Pogliano, K., Gross, H., Raaijmakers, J.M., Moore, B.S., Laskin, J., Bandeira, N., Dorrestein, P.C., 2012. Mass spectral molecular networking of living microbial colonies. Proceedings of the National Academy of Sciences 109, E1743-E1752.

Wolf, S., Schmidt, S., Muller-Hannemann, M., Neumann, S., 2010. In silico fragmentation for computer assisted identification of metabolite mass spectra. BMC Bioinformatics 11, 148.

Wolfersberger, M.G., Pieringer, R.A., 1974. Metabolism of sulfoquinovosyl diglyceride in *Chlorella pyrenoidosa* by sulfoquinovosyl monoglyceride:fatty acyl CoA acyltransferase and sulfoquinovosyl glyceride:fatty acyl ester hydrolase pathways. Journal of Lipid Research 15, 1-10.

Yagi, T., Benson, A.A., 1962. Plant sulfolipid. 5. Lysosulfolipid formation Biochimica et Biophysica Acta 57, 601-&.

Yang, J.Y., Sanchez, L.M., Rath, C.M., Liu, X., Boudreau, P.D., Bruns, N., Glukhov, E., Wodtke, A., de Felicio, R., Fenner, A., Wong, W.R., Linington, R.G., Zhang, L., Debonsi, H.M., Gerwick, W.H., Dorrestein, P.C., 2013. Molecular networking as a dereplication strategy. Journal of Natural Products 76, 1686-1699.

Table 1. Details on the lyso-sulfolipids putatively annotated in intracellular metabolites from *T. pseudonana*. All of the metabolites have the sulfoquinovosyl head group, and the table provides the details on the non-polar tail (fatty acid or sterol), elemental formula, exact mass, measured *m*/*z*, retention time (RT), and ChemSpider identification number for each metabolite. The set of fragments used to putatively annotate these metabolites are given in Table S1.

| **Non-polar tail** | **Elemental formula** | **Expected charged mass ([M-H]$^-$)** | **Measured *m*/*z*** | **RT (min)** | **ChemSpider #** |
|---|---|---|---|---|---|
| C14:0 | $C_{23}H_{44}O_{11}S$ | 527.253161 | 527.253276 | 28.5 | 8134199 |
| C16:0 | $C_{25}H_{48}O_{11}S$ | 555.284472 | 555.284665 | 31.6 | 10481089 |
| C16:1 | $C_{25}H_{46}O_{11}S$ | 553.268786 | 553.268738 | 29.4 | 8113163 |
| Sterol | $C_{27}H_{44}O_{11}S$ | 575.253161 | 575.253434 | 27.2 | 9672866 |

Table 2. Metadata associated with the additional samples containing the C14:0 lyso-sulfolipid. The table includes a brief description of each set of samples, the number of samples with the C14:0 lyso-sulfolipid, and details on how the peak areas were normalized for each set of samples. All of the filters and filtrates from these studies were processed using the methods described in Kido Soule et al. (2015).

| **Label** | **Description** | **Geographic region** | **# of samples** | **Peak area normalized by:** | **Citation** |
|---|---|---|---|---|---|
| *Synechococcus elongatus* | laboratory experiment | (not applicable) | n = 12 (filters)<br>n = 10 (filtrate) | abundance (cells $ml^{-1}$) | (Fiore et al., 2015) |
| Incubation: 70 m | experiment with seawater collected from 70 m | 10º North, 55º W | n = 15 (filters)<br>n = 15 (filtrate) | concentration of total organic carbon (μM) | (unpublished) |
| Incubation: 700 m | experiment with seawater collected from 700 m | 0º North, 34º W | n = 6 (filters)<br>n = 6 (filtrate) | concentration of total organic carbon (μM) | (unpublished) |
| Net traps | net traps deployed for 24 hours at 150 m | 0º N, 34º W<br>6ºN, 41ºW<br>6ºN, 45ºW<br>7ºN, 48ºW<br>8ºN, 50ºW<br>10ºN, 55ºW | n = 6 (filters) | wet weight of filter (g) | (unpublished) |

## Figure legends

Figure 1. Schematic summarizing the analysis of the untargeted mass spectrometry data using a combination of molecular networking and MetFrag.

Figure 2. The output from molecular networking as visualized using the Cytoscape network visualization tool. Each node in the figure is an mzRT feature with MS2 fragmentation spectra, the color of the nodes indicates the experimental conditions under which the mzRT feature was found. The nodes are connected by edges and the thickness of the line is a measure of the similarity between each pair of nodes. Letters (A–M) are used to label each cluster.

Figure 3. MS2 fragmentation spectra from the four metabolites with the sulfoquinovosyl head group. The compounds differ in the non-polar tail with three of the compounds having a single fatty acid (A) C14:0, (B) C16:0, (C) C16:1, and (D) one compound with a 21-carbon sterol. The inset in (A) is sulfoquinovose, the head group for each lipid. The structures corresponding to each metabolite are given in Figure S2. The numbers in each subplot are the nominal masses for the top six MS2 fragments.

Figure 4. Three lyso-sulfo lipids with a single fatty acid were putatively annotated in the experiment with *T. pseudonana*. The lipids contained (A) a 14:0 fatty acid, (B) a 16:0 fatty acid, or (C) a 16:1 fatty acid and all of them showed higher cell-specific levels at the conclusion of the experiment when *T. pseudonana* was grown under phosphate-limited conditions.

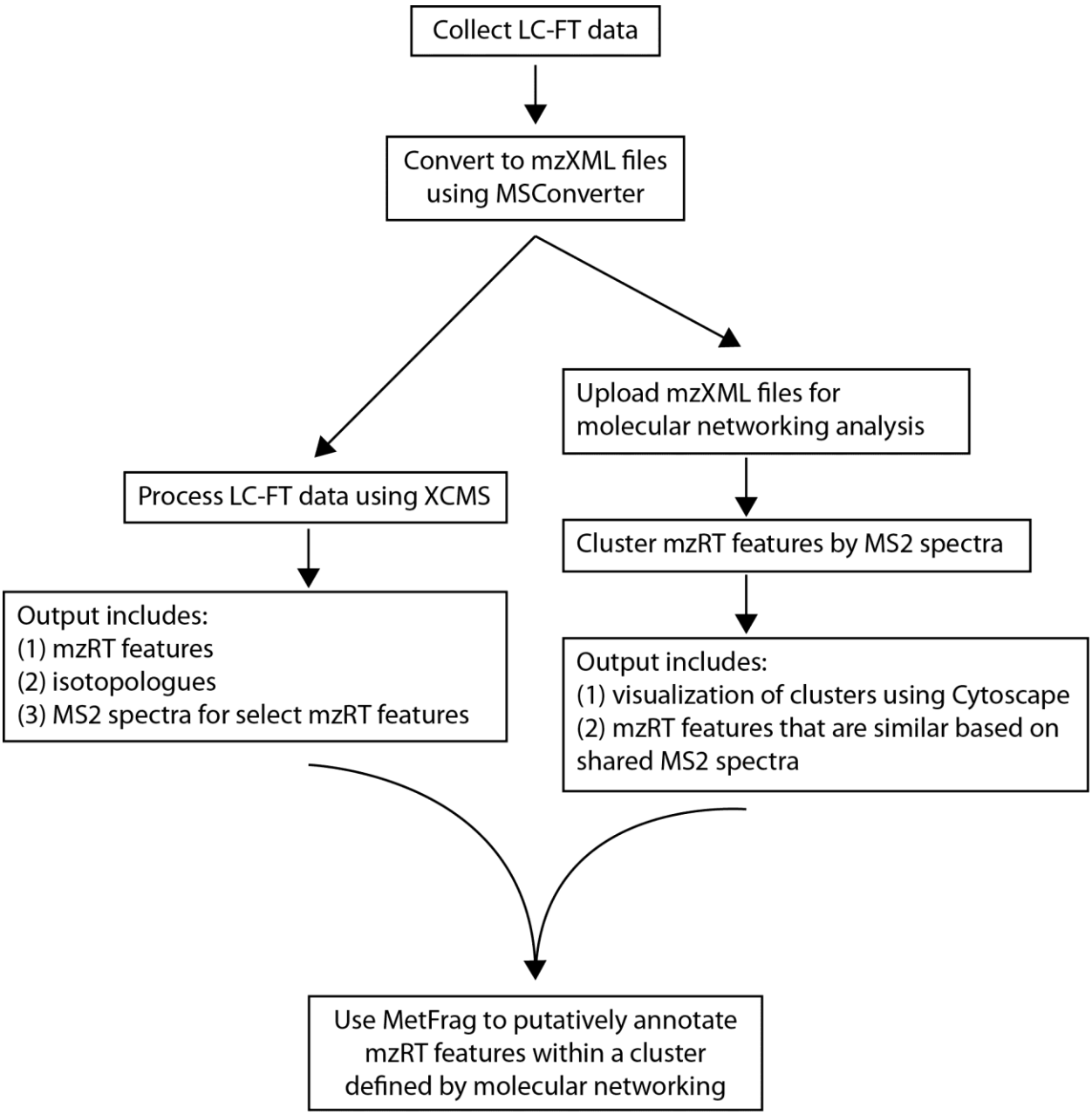
Figure 5. A sterol with the sulfoquinovosyl head group showed generally higher cell-specific levels under the phosphate-limited growth conditions. The box represents the middle 50% of

500 the data, or the inter-quartile range (IQR); whiskers extending above and below the box include
501 data within 1.5 IQRs of the box. The lines in boxes are median values.

502 Figure 6. The lyso-sulfo lipid with the C14:0 fatty acid was found in extracts from filters and
503 filtrates from four additional sets of samples processed within our laboratory: (A) a cultured
504 autotrophic microorganism, *S. elongatus* (Fiore et al., 2015), incubation experiments conducted
505 with seawater from (B) 70 m, and (C) 700 m, and (D) in particulate material captured by net
506 traps deployed at 150 m in the western equatorial Atlantic Ocean. The box represents the
507 middle 50% of the data, or the inter-quartile range (IQR); whiskers extending above and below
508 the box include data within 1.5 IQRs of the box; +: outliers, defined as normalized peak areas
509 between 1.5 and 3 IQRs distant from the box. The lines in boxes are median values.

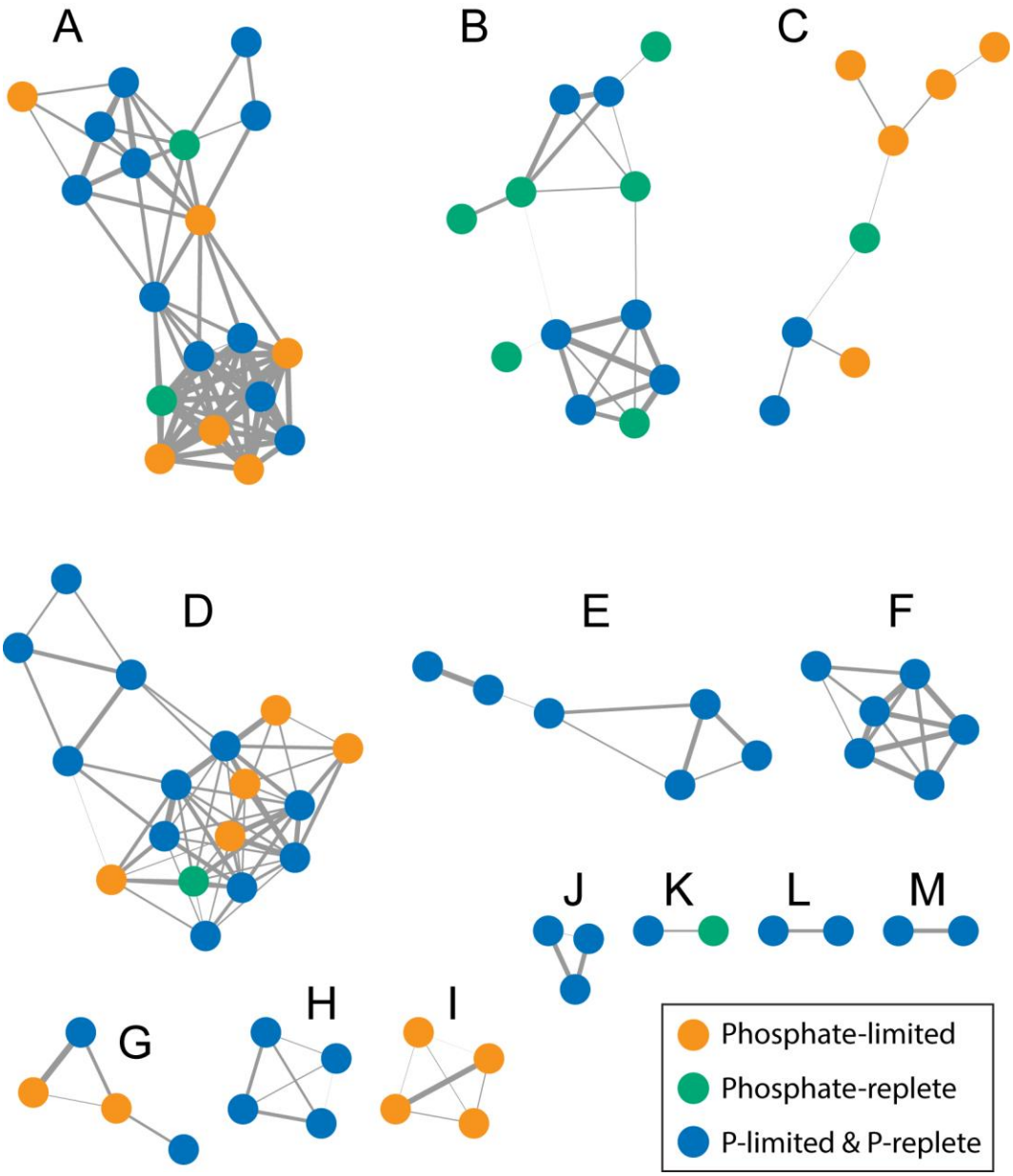
510

512 Figure 1




513

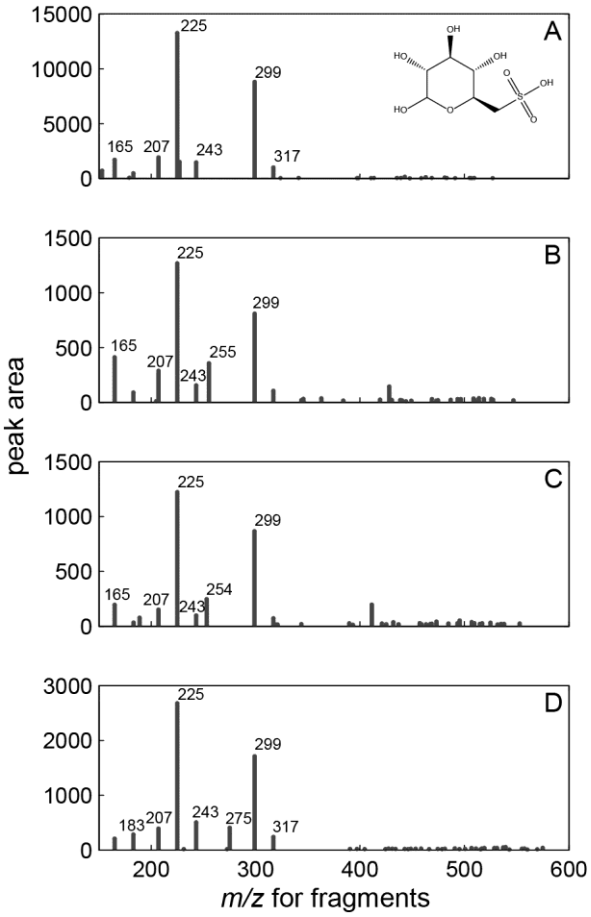
515 Figure 2




516

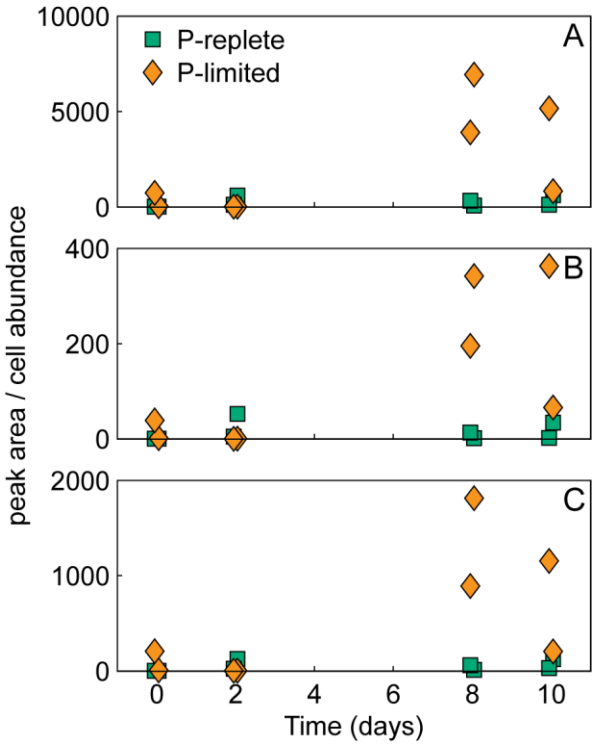
517

519 Figure 3



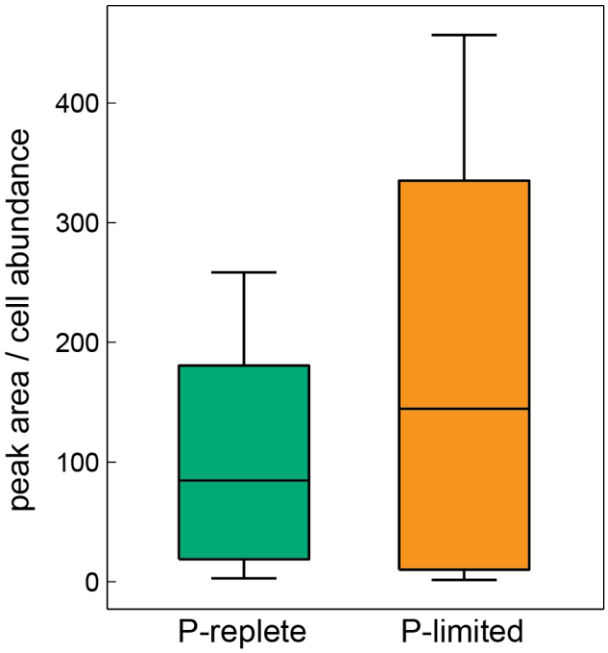

520

522 Figure 4



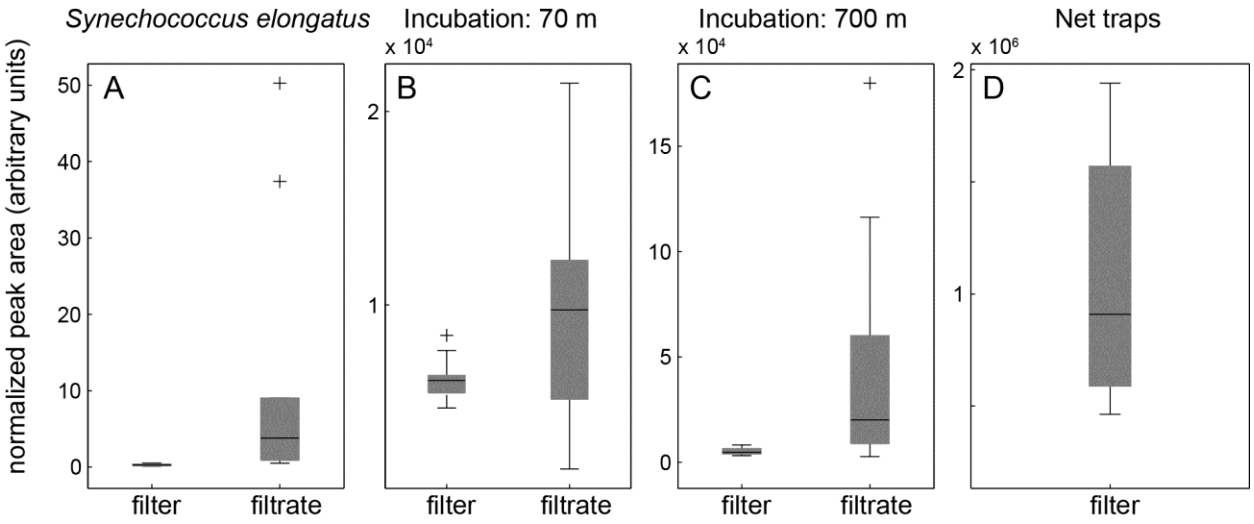

523

525 Figure 5




526

528 Figure 6




529